

# hMDAP: A Hybrid Framework for Multi-paradigm Data Analytical Processing on Spark

Xiaowang Zhang  
School of Computer Science  
and Technology  
Tianjin University, P. R. China  
Tianjin Key Laboratory of  
Cognitive Computing and  
Application  
Tianjin 300350, P.R. China  
xiaowangzhang@tju.edu.cn

Jiahui Zhang  
School of Computer Science  
and Technology  
Tianjin University, P. R. China  
Tianjin Key Laboratory of  
Cognitive Computing and  
Application  
Tianjin 300350, P.R. China  
zhangjiahui@tju.edu.cn

Zhiyong Feng  
School of Computer Software  
Tianjin University, P. R. China  
Tianjin Key Laboratory of  
Cognitive Computing and  
Application  
Tianjin 300350, P.R. China  
zyfeng@tju.edu.cn

## ABSTRACT

We propose hMDAP, a hybrid framework for large-scale data analytical processing on Spark, to support multi-paradigm process (incl. OLAP, machine learning, and graph analysis etc.) in distributed environments. The framework features a three-layer data process module and a business process module which controls the former. We will demonstrate the strength of hMDAP by using traffic scenarios in a real world.

## Keywords

Data analytical processing; OLAP; multi-paradigm; Spark

## 1. INTRODUCTION

Data analysis has become a useful technique to organize, process, and analyze large amounts of data in order to obtain useful knowledge effectively such as hidden patterns, implicit correlations, future trends, customer preferences, valuable business information etc [3]. OLAP (*online analytical processing*) [6], as a key technology to provide rapid access to data (mostly relational data) for analysis via multidimensional structures, enables users (e.g., analysts, managers, executives etc.) to gain useful knowledge from data in a fast, consistent, interactive accessing way. There are many popular enterprise database management systems for supporting OLAP. For example, Oracle OLAP [8, 18] is Oracle's current computing engine for online analytical processing. IBM company based on the DB2 database proposes the IBM DB2 OLAP Server [4, 2] which can analyze the relational database quickly and directly. Microsoft also provides SQL Server Analytic Services (SSAS) [19, 13] supporting for OLAP to analyze information, tables, and files scattered across multiple databases.

The characteristics of big data is not confined to only volume and velocity; it is also referred by the variety, variability and complexity of the data [11, 7]. Due to the volume, variety and velocity at which the data grows, it is extremely difficult for organisations

to process this data for timely and accurate decisions [1]. For this challenge, big data analysis [16] has become a tool to solve the problem. The primary goal of big data analysis is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence programs [16]. Recently, many techniques have been successfully developed for providing big data analysis in various applications. For example, Oracle Bigdata [14] builds on Hadoop [10] through Oracle Direct Connector connecting Hadoop and Oracle databases. SQL Server 2012 [17] provides the extension service of OLAP and business intelligence on Hadoop to support big data analysis. IBM SmartCloud provides a Hadoop-based analytical software InfoSphere BigInsights [20] which can connect with IBM DB2. However, those existing techniques of big data analysis are mostly based on OLAP which is not effective to process data in various models (e.g., semi-structure [16]), they do not always bring highly accurate analysis due to the variety and variability of big data in a complicated application—for example, the real-time data on the performance of traffic applications or of mobile applications. Besides, how to process big data analysis efficiently is always an important problem when the scale of big data grows exponentially [5].

In this demonstration, we propose a hybrid framework for big data analysis on Apache Spark [12] (a high-performance computing architecture) which builds on HDFS of Hadoop. The framework features a three-layer data process module and a business process module which controls the former. Within this framework, we can support multi-paradigm data process (i.e., a technical connectivity between various disparate process [21]) in order to improve the accuracy of analysis, where various big data analysis techniques (incl. OLAP, machine learning, and graph analysis etc.) are inter-operated to process the analysis of various applications of big data (incl. data cube [9], intelligent prediction, and complex network etc.) respectively. Moreover, our proposed framework built on Spark can process large-scale data efficiently. Finally, we implement hMDAP and demonstrate the strength of hMDAP by using traffic scenarios in a real world.

## 2. ARCHITECTURE

In Figure 1, we depict the architecture of our framework consisted of four parts: *the storage management, the resource scheduling, the query analysis and the business process*. In the following sections, we will introduce each part in detail.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

XXX XXX

© 2017 ACM. ISBN 978-1-4503-2138-9.

DOI: XX.XXX/XXXX

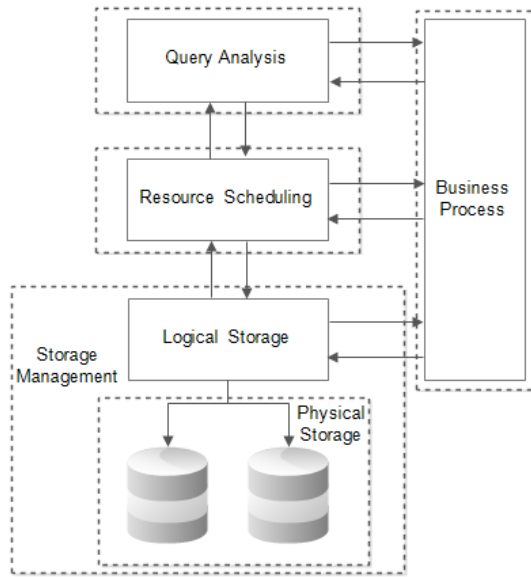


Figure 1: The hMDAP architecture.

## 2.1 Storage management

In Figure 2, there are two parts, the physical storage and the logical storage. The rapid growth of data makes the physical storage of data from single source storage to distributed storage. In order to solve the storage of multi-source data, we adopt the existing distributed file system. In our framework, it is HDFS (Hadoop Distributed File System [10]).

Besides, it products many types of data due to the different needs of applications, such as tables, texts, RCFile(the file type of Hive) and sequence data. In order to use these different types of data, we compose the abstract relational views by designing the metadata with semantics to convert data types to the relational data we can handle.

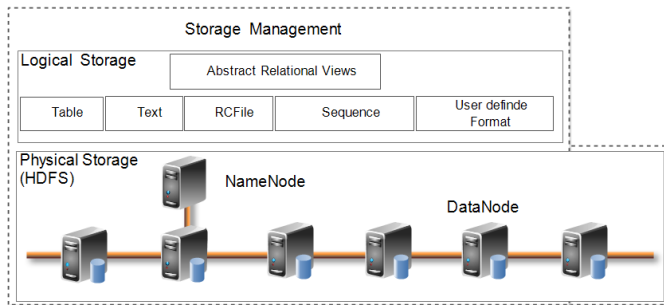


Figure 2: Storage management.

## 2.2 Resource scheduling

In our framework, the development is based on Spark and the module of the resource scheduling is assigned to Spark. The Figure 3 depicts the resource scheduling in our framework. We use MySQL [?] to query over relational database. The part of MLlib is Spark machine learning library. We call the functions in the library to compute. GraphX is the graph query module of Spark. We use it to query graphs and it provides a possibility to transform the different data formats to graph to query.

## 2.3 Query analysis

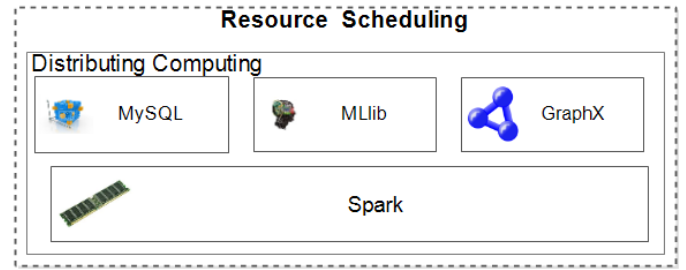


Figure 3: Resource scheduling.

The module of the query and analysis is located on the top of the framework. It is not only the entrance to provide services, but also provides the standard syntax and semantic specification of multi-paradigm data analytical processing. At present, HiveQL is similar to the standard SQL, which is oriented to the classic OLAP task, and does not deal with the query language based on ML analysis and graph data analysis. On the basis of not changing the existing query language syntax standard, we develop a multi paradigm for large data fusion analysis query language expanded of machine learning(ML) and graph analysis.

Our big data analysis and processing of the query language is based on the improvement of the fusion of SQL and HiveQL in multi-paradigm. First of all, we analyze the support of HiveQL and SQL respectively and count the amount of operations which can be supported by the traditional relational algebra model. On the basis of the relational algebra model, we add other necessary operators to construct an extension of the algebraic language model, which can fully support the operation of HiveQL and standard SQL. For the operator with higher complexity, it is split into smaller sub operator or used other methods to optimize it. For the ML analysis, we count the commonly used analytical processing methods, such as classification and clustering, and define the abstract interfaces for common ML analysis processing methods. For the graph analysis processing, we also count the commonly used analytical processing methods, such as the shortest path algorithm, and define the abstract interfaces for them.

In this module, the framework also relates to the implementation of the OLAP on the relational database and ML and graph data processing tasks on the distributed framework. The traditional relational database query optimization method is no longer applicable to this situation. According to the different characteristics of relational storage management query engine and distributed file system of computing engine, we summarize the query information and optimize the performance. Firstly, we investigate the statistical index system used in traditional database and analyze the interaction between each index and the index in the system. Then, for each index in the index system of statistical information, we design efficient and accurate sampling methods to calculate the cost model in query optimization. According to the above statistics, we can also design a storage and maintenance programs which is easy to update and manage. And we may use the cost model in the traditional relational database to design a new cost model which can reflect the query cost of the mixed data.

Figure 4 displays the query analysis. The main architectural components of the query analysis are *Query* and *Data Analysis Process Tools(DAP Tools)*. In the first part, we can query by SQL or the function user defined as specified format. The DAP tools contain classical OLAP, DAP on machine learning and DAP on graph.

## 2.4 Business process

Our framework provides an analysis method for the large scale data analysis process. But in the face of complex business pro-

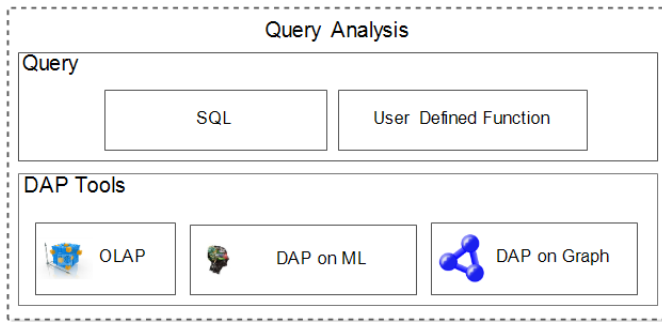


Figure 4: Query analysis.

cesses in different fields, we need the domain knowledge and according to the domain knowledge, we can design the multi-paradigm fusion of analysis task. We can draw lessons from the method of service composition in service oriented architecture design.

In this module, we need to do two things: developing a multi paradigm fusion analysis process orchestration language syntax and the complex business process scheduling method. In the first part, we need to analyze the patterns and characteristics of service orchestration language in service oriented architecture design and design an abstract model of the executable process. On the basis of the abstract model, we summarize the basic activities of complex business process analysis. Finally, we define the grammar of the business process. In the semantic, we need to research and analysis the meanings of basic business activities and define the start point, end point and the basic command. In the second part, we need to study and analyze complex business processes in practical applications. Then, we build complex business process models and refine the way to exchange messages in public business processes. After that, we need to control the interaction of each part of the resources through the interaction sequence of messages, achieving a reasonable call for each resource service. We still need to investigate the applicability of existing object-oriented design patterns. For the analysis of complex business process integration model, we design data business processes. We refine the design patterns in complex business processes based on the advantages and principles of existing design patterns.

In the real world, the business process model is complex and it takes a lot of time to analyze. The Figure 5 illustrates the details of the business process in our framework.

The user needs to write the configuration files before he or she submits the query. The format of configuration files are shown in Section 3. When the user submits a query to the framework, the query and analysis module in the framework starts to parse the user's query. This module parses queries according to predefined semantics, such as XML(Extensible Markup Language). The module transforms the user's query to two parts, the query over relational databases and the query in machine learning. We default that the user's query including the query over relational databases and the module determines whether or not to carry out the query in machine learning. We think that when the result of the query over relational databases is null, the framework begins to query in machine learning. After the analysis module, the framework uses the query over relational databases and the information about the databases which is read from the configuration files to query the relational databases. Then, the framework runs the query in machine learning. The input of machine learning is the result of querying by relational databases which the query statement is stored in the configuration files. And the parameters of the machine learning algorithm is also stored in the configuration files. When the framework gets the information of the machine learning algorithm, it starts to

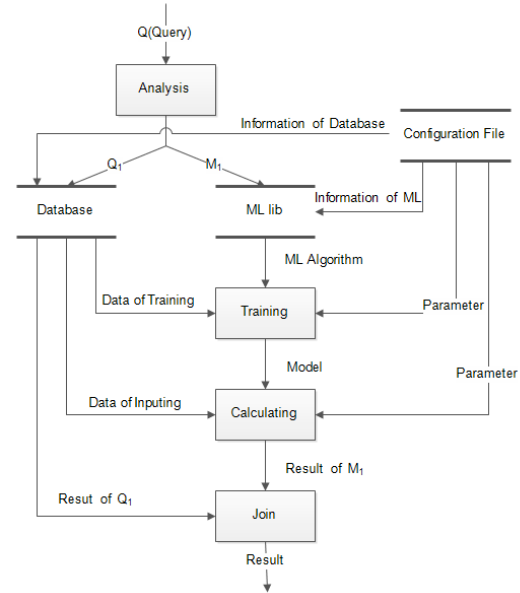


Figure 5: Business process.

Figure 6: Query interface of hMDAP.

train and calculate and the parameters of the training of the machine learning also comes from the configuration files. Finally, the framework makes a join of the results of two parts.

### 3. DEMONSTRATION

In this section, we present the interface of hMDAP based in Javascript, which communicate with the service in Java. We show the screenshot of hMDAP in Figure 6 and the configuration file we mentioned above in Figure 7.

The interface is composed as follows:

- Configuration of Machine Learning: it is a text to input the path of the configuration file of the machine learning algorithm, such as parameters.
- Configuration of Relation Database: it is a text to input the path of the configuration file of the relational databases, such as the user name.
- Results: it is a text to display the results of the background.
- Run: it is a button to start the program and when the program runs over, the results are shown in the *Results*.
- Save: it is a button to save the context from *Results* in text file and at the same time, empty all the text box contents.
- Cancel: cancel the running of this program and empty all the text box contents.

```

<configuration>
  <input>
    <database>
      <url>jdbc:mysql://127.0.0.1:3306/traffic</url>
      <user>root</user>
      <password>rootPassword</password>
      <sql>select driverId,lon,lat from gps</sql>
    </database>
  </input>
  <parameter>
    <value>3</value>
    <value>10</value>
    <value>2</value>
  </parameter>
  <algorithm>KMeans</algorithm>
</configuration>

```

**Figure 7: The configuration file of the machine learning.**

And the details of the configuration file is as follows:

- configuration: it is the beginning of the configuration file.
- input: it is the training dataset of the machine learning algorithm.
- database: it indicates that the input dataset comes from the relational database as following information
- url, user, password: they are the parameters to connect to the relational database, the location of the database, the user name and the password of the user.
- sql: it is the statement to query the relational database.
- parameter: the contents under this label are the parameters of the machine learning algorithms except the input parameter.
- value: a series of these labels are the values of the parameters.
- algorithm: it is the name of algorithm. For example, the value of *algorithm* is *KMeans* and our framework runs the algorithm named *KMeans* which is defined in our library. User can customize the algorithm and give the location of the algorithm in this label.

Before running the interface, the user should write two configuration files, the configuration of machine learning algorithms as Figure 7 and the configuration of relational databases that the contexts are the parts of *<database>* in Figure 7. When the user writes two files, he or she should write the paths of the files in the texts on the interface. Then, click the button *Run*. If the user wants to save the results, he or she clicks the button of *Save*. If the user don't need the results, he or she clicks the button of *Cancel*.

## 4. CONCLUSION

In this demonstration, we proposed hMDAP, a hybrid framework for large-scale data analytical processing to support multi-paradigm process on Spark. The multi-paradigm processing mechanism of hMDAP can provide the interoperability of data analytical process techniques to process data which might be not effectively handled if we only apply single data analytical process technique. On the other hand, hMDAP takes advantage of the high-performance of Spark in processing large-scale data. We believe that hMDAP provides a new approach to big data analysis in a multi-paradigm way.

## Acknowledgments

This work is supported by the programs of the Key Technology Research and Development Program of Tianjin (16YFZCGX00210), the National Key Research and Development Program of China (2016YFB1000603), the National Natural Science Foundation of China (NSFC) (61672377), and the Open Project of Key Laboratory of Computer Network and Information Integration, Ministry of Education (K93-9-2016-05). Xiaowang Zhang is supported by Tianjin Thousand Young Talents Program.

## 5. REFERENCES

- [1] Ahmed H. (2015). Importance of big data analytics for business growth. : *BIG Data Analytics News*

- http://bigdataanalyticsnews.com/importance-of-big-data-analytics-for-business-growth/
- [2] Baragoin C., Bercianos J., Komel J., Robinson G., Sawa R., and Schuinder E. (2001). DB2 OLAP server theory and practices. *International Technical Support Organization*.
- [3] Berson A. and J. Smith S. (1997). Data warehousing, data mining, and OLAP. *McGraw-Hill*.
- [4] Bontempo C. and Zagelow G.(1998). The IBM data warehouse architecture. *Commun. ACM*, 1998, 41(9): 38-48.
- [5] Campbell P. (editor). (2008). Big data: science in the petabyte era. *Nature*, 455:1â&#136.
- [6] Chaudhuri S. and Dayal U. (1997). An overview of data warehousing and OLAP technology. *SIGMOD Record*, 26(1): 65-74.
- [7] Chen H., Chiang R.H.L., and Storey V.C. (2012). Business intelligence and analytics: From big data to big Impact. *MIS quarterl*, 36(4):1165-1188.
- [8] Dodge G. and Gorman T.(1998). Oracle data warehousing. John Wiley & Sons, Inc.
- [9] Gray J., Chaudhuri S., Bosworth A., Layman A., Reichart D., Venkatrao M., Pellow F., and Pirahesh H. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Min. Knowl. Discov.*, 1(1): 29-53.
- [10] Hadoop. (2015). <http://hadoop.apache.org/>
- [11] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and H. Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.
- [12] Meng X., K.Bradley J., Yavuz B., R.Sparks E., Venkataraman S., Liu D., Freeman J., B.Tsai D., Amde M., Owen S., Xin D., Xin R., M. Franklin M., Zadeh Z., Zaharia M., and Talwalkar A. (2016). MLlib: Machine learning in Apache Spark. *J. Mach. Learn. Res.*, 17:1-7.
- [13] Microsoft Analysis services. (2016). SQL server 2016 and later. <https://msdn.microsoft.com/en-us/library/bb522607.aspx>
- [14] Plunkett T., Macdonald B., Nelson B., HornickM., Sun H., Mohiuddin K., Harding D., Mishra G., Stackowiak R., Laker, K., and Segleau D. (2013). Oracle big data handbook. *McGraw-Hill Osborne Media*
- [15] P. Sheth A., Kochut K., A. Miller J., Worah D., Das S., Lin C., Palaniswami D., Lynch J., and Shevchenko I. (1996). Supporting state-wide immunisation tracking using multi-paradigm workflow technology. In: *Proc. of VLDB'96*, pp. 263-273.
- [16] Rouse W. (2012). What is big data analytics? *TechTarget.com* <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>
- [17] Sarkar D. (2013). Microsoft SQL server 2012 with Hadoop. *Packt Publishing*
- [18] Schrader M. and Vlamis D.(2009). Oracle Essbase & Oracle OLAP. *Peter Gbolagade Akintunde*.
- [19] Spofford G. (2001). MDX solutions: with Microsoft SQL server analysis services. *Wiley*.
- [20] Zikopoulos P. (2011). Understanding big data: Analytics for enterprise class Hadoop and streaming data. *McGraw-Hill Osborne Media*
- [21] zur Muehlen M. and Rosemann M.(2004). Multi-paradigm process management. In: *Proc. of CAiSE'04*, pp. 169-175.